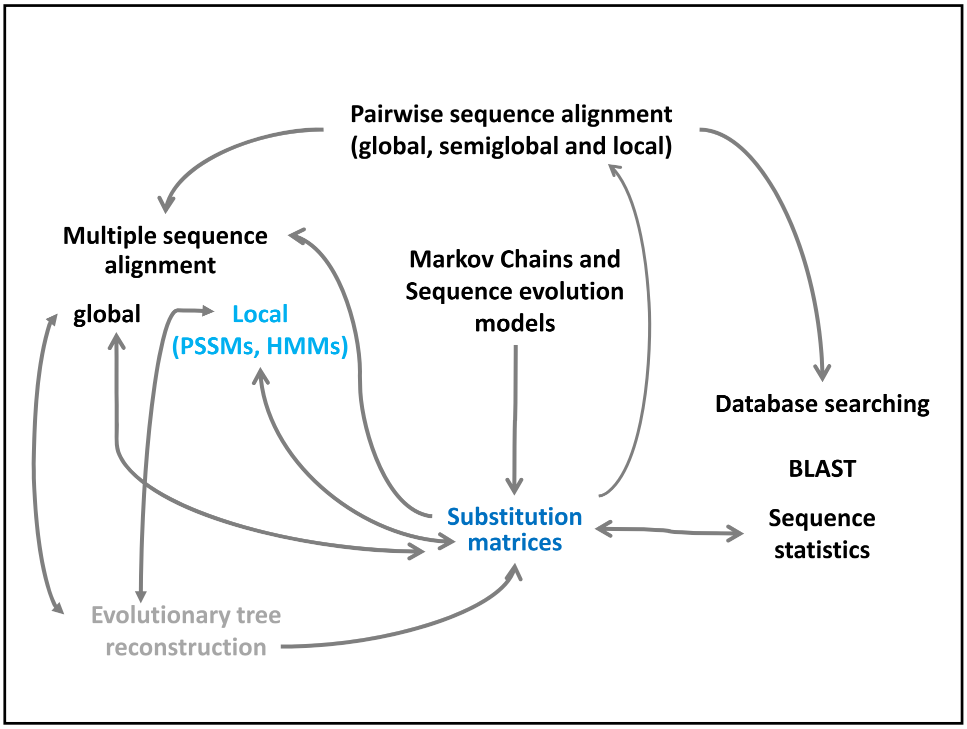


Pairwise sequence alignment (global, semiglobal and local)

Multiple sequence alignment

global

Local (PSSMs, HMMs)

Markov Chains and Sequence evolution models

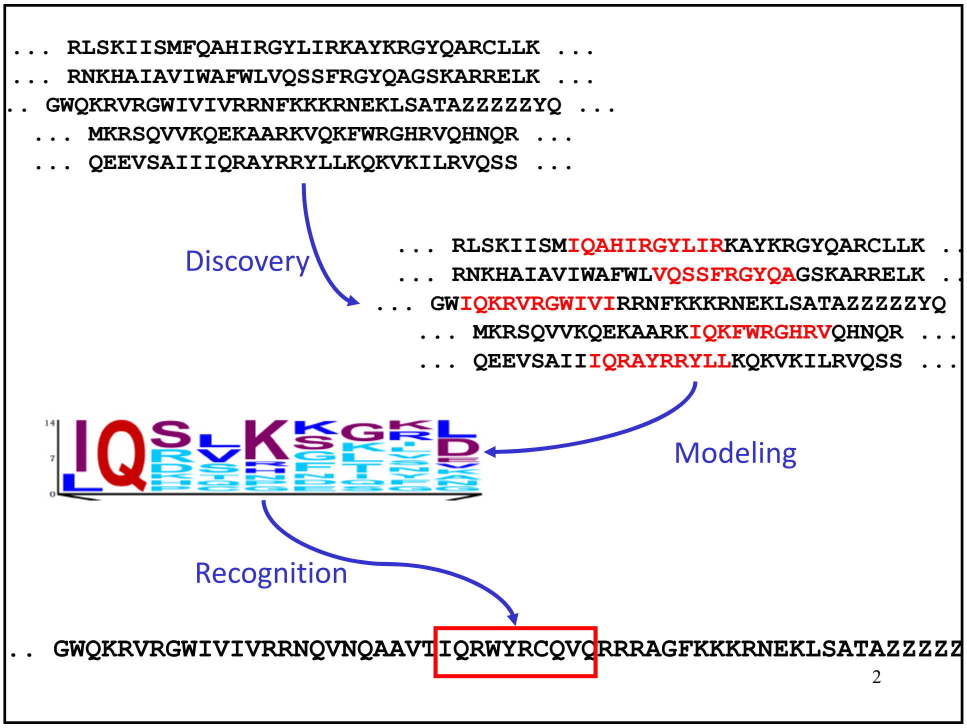
Database searching

BLAST

Substitution matrices

Sequence statistics

Evolutionary tree reconstruction



```
... RLSKIISMFQAHIRGYLIRKAYKRGYQARCLLK ...
... RNKHAIAVIWAFWLVQSSFRGYQAGSKARRELK ...
.. GWQKRVRGWIVIVRRNFKKKRNEKLSATAZZZZZYQ ...
  ... MKRSQVVKQEKAARKVQKFWRGHRVQHNQR ...
  ... QEEVSAIIIQRAYRRYLLKQKVKILRVQSS ...
```

Discovery

```
... RLSKIISMIQAHIRGYLIRKAYKRGYQARCLLK ..
... RNKHAIAVIWAFWLVQSSFRGYQAGSKARRELK ..
... GWIQKRVRGWIVIRRNFKKKRNEKLSATAZZZZZYQ
  ... MKRSQVVKQEKAARKIQKFWRGHRVQHNQR ...
  ... QEEVSAIIIQRAYRRYLLKQKVKILRVQSS ...
```

Modeling

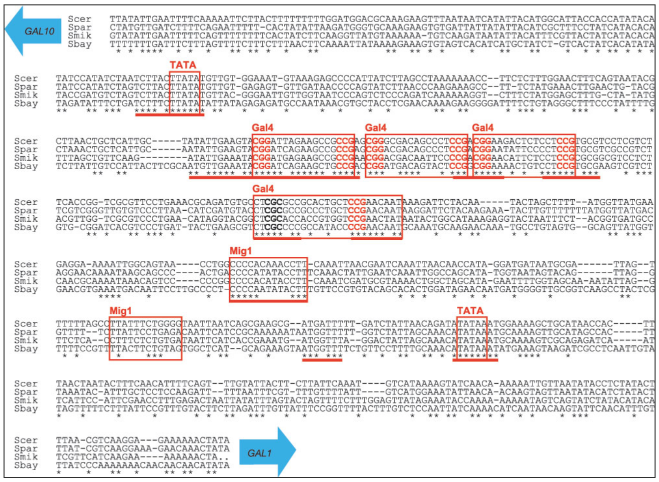
Recognition

```
.. GWQKRVRGWIVIVRRNQVNQAAVTIQRWYRCQVQRRRAGFKKKRNEKLSATAZZZZ
```

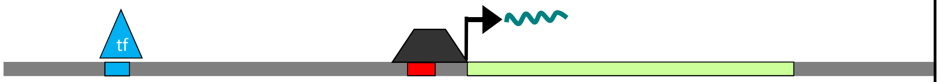
2

Conserved patterns in biological sequences

Example: Transcription factor binding sites



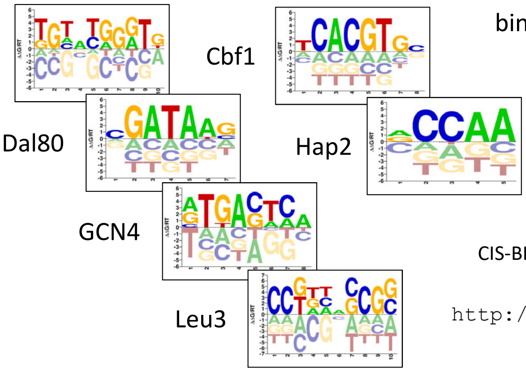
Kellis *et al, Nature*, 03



transcription factor binding site

promotor

Transcription factor binding site motifs



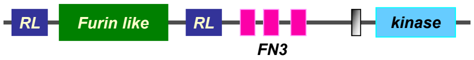
CIS-BP, Catalog of Inferred Sequence Binding Preferences

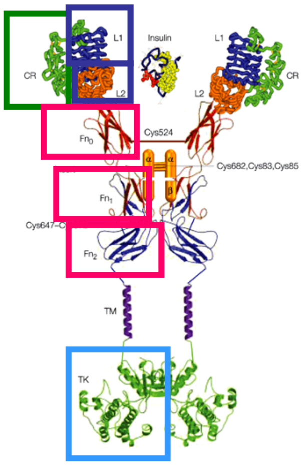
`http://cisbp2.ccbr.utoronto.ca`

Conserved patterns in biological sequences

Example: Protein domains

- Fold independently
- Carry out specific functions
- Found in diverse contexts
- Conserved in evolution

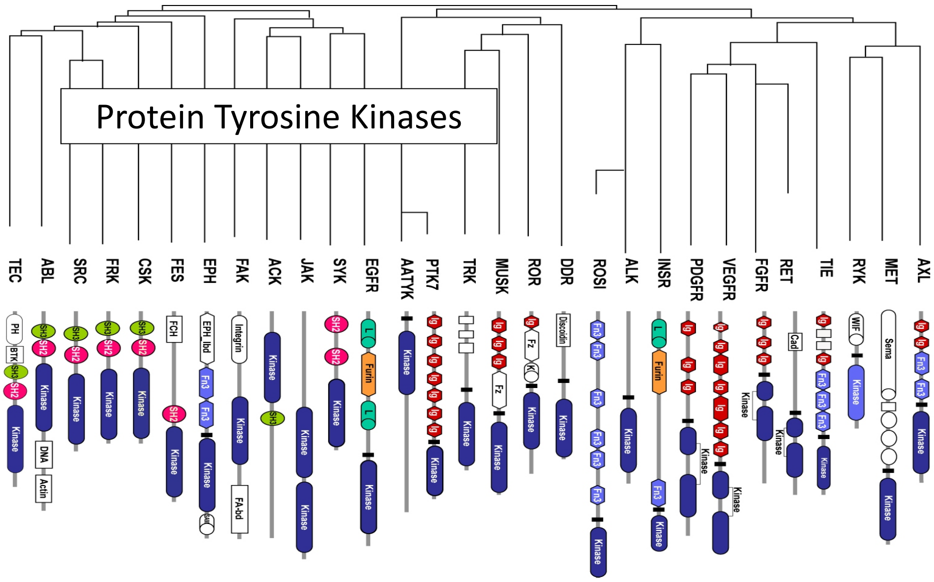




Insulin receptor

5

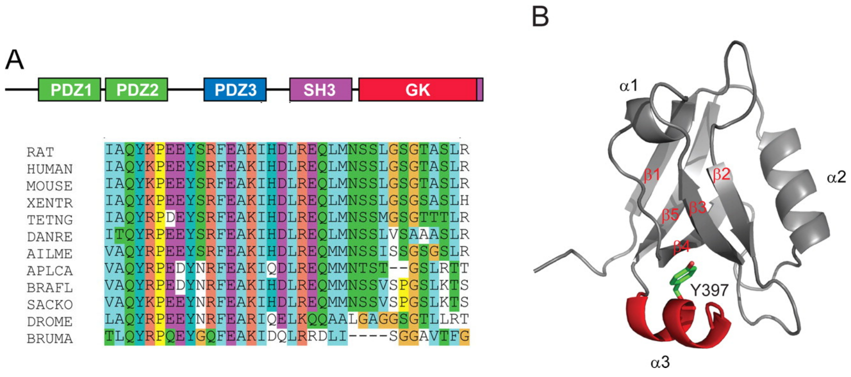
Protein Tyrosine Kinases



Adapted from Robinson *et al*., 2000

6

Domain architecture of PSD-95 and crystal structure of PDZ3.

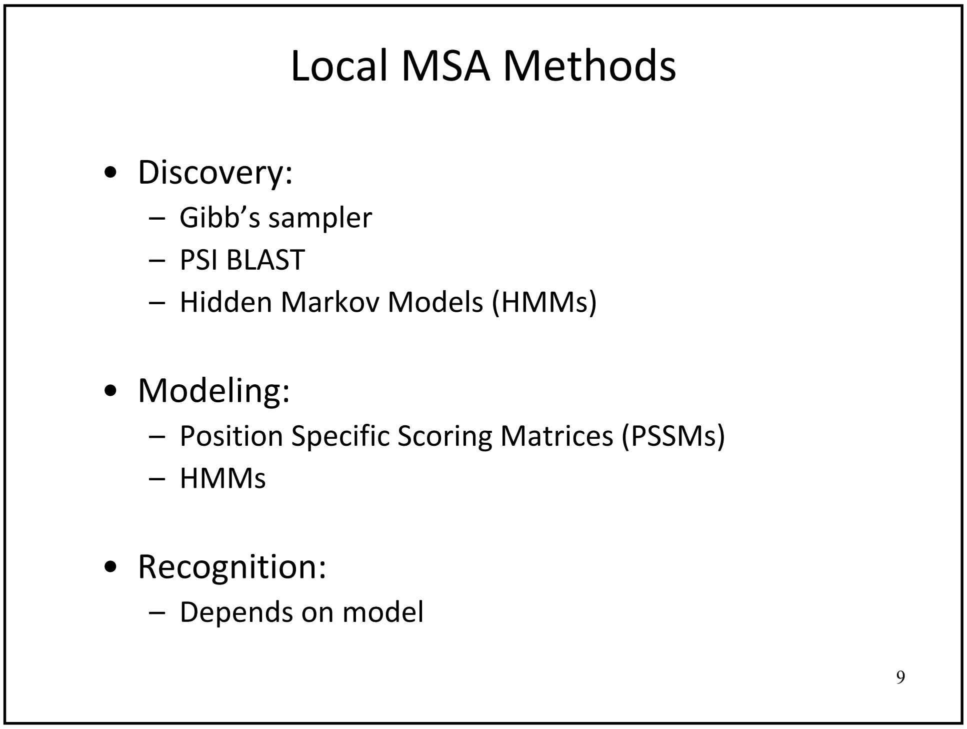


```
RAT    IAQYKPEEYSRFEAKIHDLREQLMNSSLGSGTASLR
HUMAN  IAQYKPEEYSRFEAKIHDLREQLMNSSLGSGTASLR
MOUSE  IAQYKPEEYSRFEAKIHDLREQLMNSSLGSGTASLR
XENTR  IAQYKPEEYSRFEAKIHDLREQLMNSSLGSGSASLH
TETNG  IAQYRPDEYSRFEAKIHDLREQLMNSSMGSGTTTLR
DANRE  ITQYRPEEYSRFEAKIHDLREQLMNSSLVSAAASLR
AILME  VAQYRPEEYSRFEAKIHDLREQMMNSSISSGSGSLR
APLCA  VAQYRPEDYNRFEAKIQDLREQMMNTST--GSLRTT
BRAFL  VAQYRPEDYNRFEAKIHDLREQMMNSSVSPGSLKTS
SACKO  VAQYKPEEYNRFEAKIHDLREQMMNSSVSPGSLKTS
DROME  LAQYRPEEYNRFEARIQELKQQAALGAGGSGTLLRT
BRUMA  TLQYRPQEYGQFEAKIDQLRRDLI----SGGAVTFG
```

Zhang J et al. J. Biol. Chem. 2011;286:41776-41785

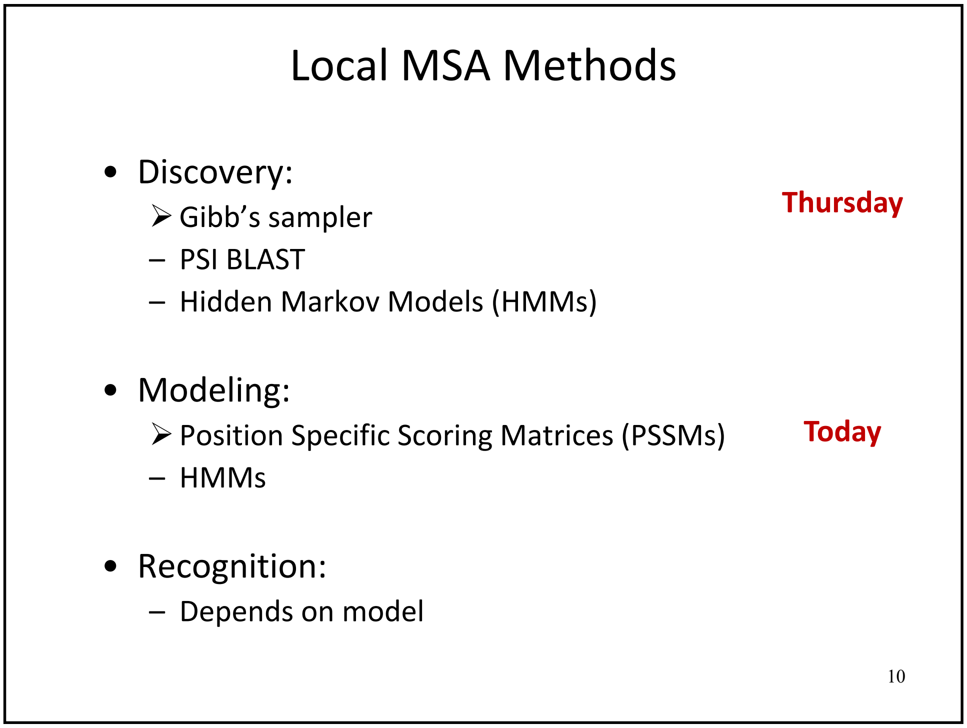
Local Multiple Sequence Alignment
Probabilistic Framework

- Discovery
  - Given multiple sequences, often unaligned, find a conserved pattern or *motif*
- Representation
  - Given an alignment of the motif (often ungapped), construct probabilistic model summarizing conserved features
- Recognition (using model)
  - Given a new sequence, does it contain the motif?
  - Find all sequences in a database that have the *motif*.



Local MSA Methods

- Discovery:
  - – Gibb’s sampler
  - – PSI BLAST
  - – Hidden Markov Models (HMMs)
- Modeling:
  - – Position Specific Scoring Matrices (PSSMs)
  - – HMMs
- Recognition:
  - – Depends on model



Local MSA Methods

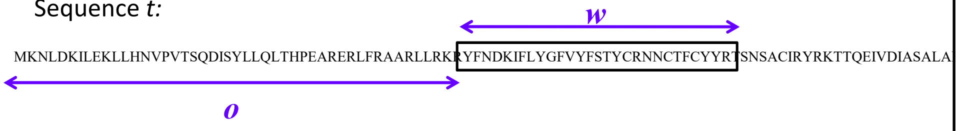
- Discovery:
  - ➢ Gibb’s sampler **Thursday**
  - – PSI BLAST
  - – Hidden Markov Models (HMMs)
- Modeling:
  - ➢ Position Specific Scoring Matrices (PSSMs) **Today**
  - – HMMs
- Recognition:
  - – Depends on model

Scoring a potential new instance of the pattern:

Given a sequence *t,* a window of length *w* starting at the next position after offset *o* is scored as follows*:*

$$Score[t,o]=\sum_{i=1}^{w}S[t[o+i],i]$$

Sequence *t:*


w
MKNLDKILEKLLHNVPVTSQDISYLLQLTHPEARERLFRAARLLRKRYFNDKIFLYGFVYFSTYCRNNCTFCYYRTSNSACIRYRKTTQEIVDIASALA
o

This score can be interpreted as a log likelihood ratio…

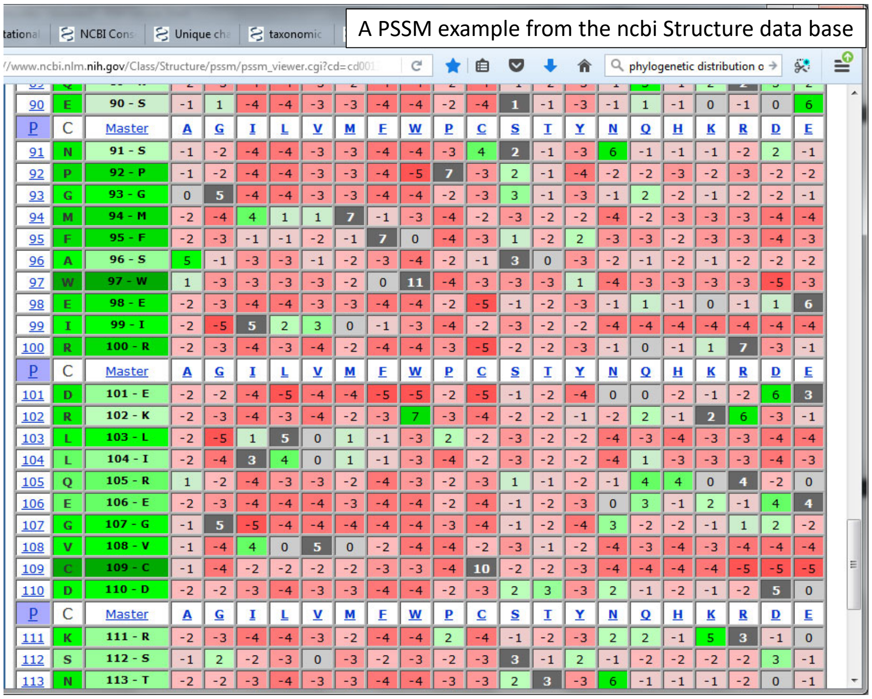
A PSSM is a log odds scoring matrix

Note that the score of a window of length *w* at position *o* in *t*, is a log likelihood ratio of the form

$$S[t,o]=\log_2\frac{P[data\,|\,H_a]}{P[data\,|\,H_0]}$$

where the *data* is the subsequence at *o, H_a* is the alternate hypothesis that *t* contains the pattern and *H_0* is the null hypothesis (no pattern, background frequencies)

$$\begin{aligned}S[t,o]&=\sum_{i=1}^{w}S[t[o+i],i]\\&=\sum_{i=1}^{w}\log_2 P[t[o+i],i]\\&=\sum_{i=1}^{w}\log_2\frac{q[t[o+i],i]}{p(t[o+i])}\\&=\log_2\frac{\prod_{i=1}^{w}q[t[o+i],i]}{\prod_{i=1}^{w}p(t[o+i])}\\&=\log_2\frac{P[data|\,H_a]}{P[data|\,H_0]}\end{aligned}$$



A PSSM example from the ncbi Structure data base

tational
NCBI Cons
Unique cha
taxonomic
/www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm_viewer.cgi?cd=cd001
phylogenetic distribution o

90	E	90 - S	-1	1	-4	-4	-3	-3	-4	-4	-2	-4	1	-1	-3	-1	1	-1	0	-1	0	6
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
91	N	91 - S	-1	-2	-4	-4	-3	-3	-4	-4	-3	4	2	-1	-3	6	-1	-1	-1	-2	2	-1
92	P	92 - P	-1	-2	-4	-4	-3	-3	-4	-5	7	-3	2	-1	-4	-2	-2	-3	-2	-3	-2	-2
93	G	93 - G	0	5	-4	-4	-3	-3	-4	-4	-2	-3	3	-1	-3	-1	2	-2	-1	-2	-2	-1
94	M	94 - M	-2	-4	4	1	1	7	-1	-3	-4	-2	-3	-2	-2	-4	-2	-3	-3	-3	-4	-4
95	F	95 - F	-2	-3	-1	-1	-2	-1	7	0	-4	-3	1	-2	2	-3	-3	-2	-3	-3	-4	-3
96	A	96 - S	5	-1	-3	-3	-1	-2	-3	-4	-2	-1	3	0	-3	-2	-1	-2	-1	-2	-2	-2
97	W	97 - W	1	-3	-3	-3	-3	-2	0	11	-4	-3	-3	-3	1	-4	-3	-3	-3	-3	-5	-3
98	E	98 - E	-2	-3	-4	-4	-3	-3	-4	-4	-2	-5	-1	-2	-3	-1	1	-1	0	-1	1	6
99	I	99 - I	-2	-5	5	2	3	0	-1	-3	-4	-2	-3	-2	-2	-4	-4	-4	-4	-4	-4	-4
100	R	100 - R	-2	-3	-4	-3	-4	-2	-4	-4	-3	-5	-2	-2	-3	-1	0	-1	1	7	-3	-1
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
101	D	101 - E	-2	-2	-4	-5	-4	-4	-5	-5	-2	-5	-1	-2	-4	0	0	-2	-1	-2	6	3
102	R	102 - K	-2	-3	-4	-3	-4	-2	-3	7	-3	-4	-2	-2	-1	-2	2	-1	2	6	-3	-1
103	L	103 - L	-2	-5	1	5	0	1	-1	-3	2	-2	-3	-2	-2	-4	-3	-4	-3	-3	-4	-4
104	L	104 - I	-2	-4	3	4	0	1	-1	-3	-4	-2	-3	-2	-2	-4	1	-3	-3	-3	-4	-3
105	Q	105 - R	1	-2	-4	-3	-3	-2	-4	-3	-2	-3	1	-1	-2	-1	4	4	0	4	-2	0
106	E	106 - E	-2	-3	-4	-4	-4	-3	-4	-4	-2	-4	-1	-2	-3	0	3	-1	2	-1	4	4
107	G	107 - G	-1	5	-5	-4	-4	-4	-4	-4	-3	-4	-1	-2	-4	3	-2	-2	-1	1	2	-2
108	V	108 - V	-1	-4	4	0	5	0	-2	-4	-4	-2	-3	-1	-2	-4	-3	-4	-3	-4	-4	-4
109	C	109 - C	-1	-4	-2	-2	-2	-2	-3	-3	-4	10	-2	-2	-3	-4	-4	-4	-4	-5	-5	-5
110	D	110 - D	-2	-2	-3	-4	-3	-3	-4	-4	-2	-3	2	3	-3	2	-1	-2	-1	-2	5	0
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
111	K	111 - R	-2	-3	-4	-4	-3	-2	-4	-4	2	-4	-1	-2	-3	2	2	-1	5	3	-1	0
112	S	112 - S	-1	2	-2	-3	0	-3	-2	-3	-2	-3	3	-1	2	-1	-2	-2	-2	-2	3	-1
113	N	113 - T	-2	-2	-3	-4	-3	-3	-4	-4	-3	-3	2	3	-3	6	-1	-1	-1	-2	0	-1

First in-class exam results

- Max: 98
- Mean: 85.1
- Median: 86
- Minimum: 66

98
97
96
95
94
93
92
91
90.5
88.5
88
86
86
85
83.5
83
81
76.5
76
73.5
70
68
66